Volume 15 Number 6 1987 Nucleic Acids Research

The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis

Gregory J.Phillips, Jonathan Arnold and Robert Ivarie

Department of Genetics, University of Georgia, Athens, GA 30602, USA

Received February 10, 1987; Accepted February 18, 1987

ABSTRACT

As shown in the accompanying paper (5), the oligonucleotide composition of the E. coli genome is highly asymmetric for sequences up to 6 bp in length when ranked from highest to lowest abundance. We show here that this largely reflects codon usage because heavily used codons were found in the highly abundant oligomers whereas rarely used codons, with some exceptions, occurred in sequences in low abundance. Furthermore, linear regression analysis revealed a strong correlation between the frequencies of each trinucleotide and its usage as a codon. Dinucleotides are also not randomly distributed across each codon position and the dinucleotide composition of genes that are transcribed but not translated (rRNA and tRNA genes) was highly related to that seen in genes encoding polypeptides. However, 45 tetra-, 8 penta-, and 6 hexanucleotides were significantly over- or underabundant by Markov chain analysis and could not be accounted for by codon usage. Of these underrepresented sequences, many were palindromes, including the Dam methylation site.

INTRODUCTION

Because the majority of Escherichia coli DNA codes for proteins, codon usage strongly influenced the observed oligonucleotide composition of the bacterial genome. However, no one has yet reported methods for measuring this effect. Also, many sequences must be reserved for other cellular processes such as transcription (promoters, terminators, operators), replication (origins), translation (initiation and termination sites), restriction/modification, recombination (chi) and repair. Many of these sequences might be underrepresented in coding DNA because their occurrence at a high frequency would be incompatible with sufficiently long coding sequences. In addition, sequences that promote disruptive DNA structures, such as palindromes (1,2) or alternating purine and pyrimidine runs which can form Z DNA (3,4) may also be underrepresented in the E. coli genome.

We have previously shown that a Markov chain is a relatively accurate predictor of the frequencies of oligonucleotides longer than 4 bases (5). Here we apply two methods to show that the overall di- through hexanucleo-

tide composition of 74,444 bp of E. coli DNA is largely set by codon usage. Nonetheless, several oligonucleotides have been identified by Markov chain analysis that were over- or underrepresented for reasons other than codon usage.

METHODS

Sequence Analysis

All E. coli DNA sequence data were taken from Genbank (Fall, 1984 update). Genes and their flanking regions were those previously described and listed (5). Two subsets of this data set were: (i) protein coding DNA of 47,358 bases containing only those sequences encoding proteins whose translational start and stop sites were known and (ii) rRNA and tRNA genes of 12,336 bases representing sequence that are transcribed but not translated. Given the size of the data sets, variations greater than 5% were statistically significant at a confidence level of 0.001 or less. Sequence editing and compiling were performed on a Digital PDP-11/34A as described (5) using a software package (DNASEQ) for analyzing DNA sequences (6). Residual values and likelihood ratio tests were performed as described (5).

RESULTS AND DISCUSSION

As shown in the accompanying paper (5), di- and trinucleotides were not well-predicted by zero and 1st order Markov chains, respectively. However, tetranucleotide frequencies were reasonably well predicted from their component di- and trinucleotides by a 2nd order Markov chain. Furthermore, accuracy improved as oligonucleotide length increased with higher order Markov chains. Hence, major variation in frequencies occurred at the di- and trinucleotide levels.

That this variation derives largely from codon usage is implied by the observation that in coding DNA mononucleotides are unequally distributed in each of the three codon positions (Table 1). These "contextual" constraints on mononucleotides appear to be influenced by the translation machinery because the base composition of codons is related to the abundance of cognate tRNA's, as noted by Ikemura (7-9). Furthermore, dinucleotides do not occur randomly at each codon position as illustrated in Figure 1. Only TT and AC approached the random 33.3% frequency across the three positions. Biases are apparent in the distributions which correlated with mononucleotide frequencies. For instance, G is highly abundant in the first position (36.7%) and infrequent in the second position (17.4%). GA

Table 1: Mononucleotide frequencies (%) at each codon position in the 47,358 bp of protein coding sequences.

	Codon Position		
	first	second	third
T	14.7	29.8	25.2
C	24.1	22.8	29.0
A	24.5	30.0	18.4
G	36.7	17.4	27.5

was the most abundant dinucleotide in XYN and one-half of the six dinucleotides occurring infrequently in the XYN position (TG, CG, TC, CC, AG, TA) contain G in the second position.

A more direct measure of the effect of codon usage can be seen in Figure 2 in which a linear regression analysis was undertaken comparing the frequencies of trinucleotides in the 74,444 bp data base to codon usage in the protein coding data base of 47,358 bp. A strong correlation was found with a coefficient of correlation of 0.649. Moreover, the complement ratios for 9 of the 10 most abundant trinucleotides were greater than one

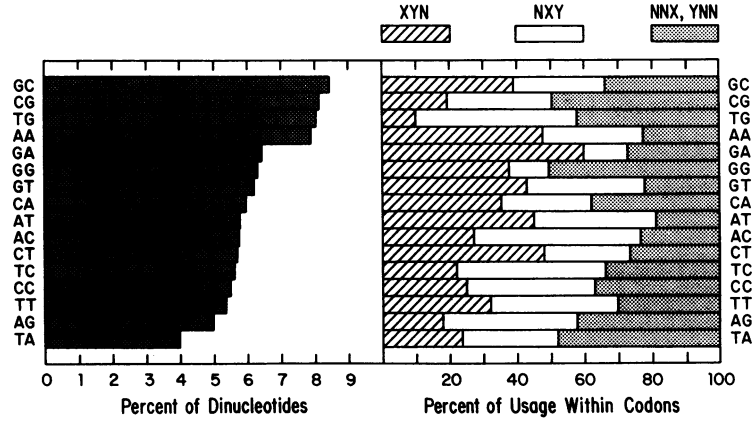

Figure 1. Ordered abundance curve of dinucleotides in E. coli coding DNA and their distribution across each codon position.

On the left, the ordered abundance of dinucleotides in 47,358 bp of coding DNA has been plotted as a histogram and in the right panel, their frequency at each codon position has been plotted after normalizing to 100%.

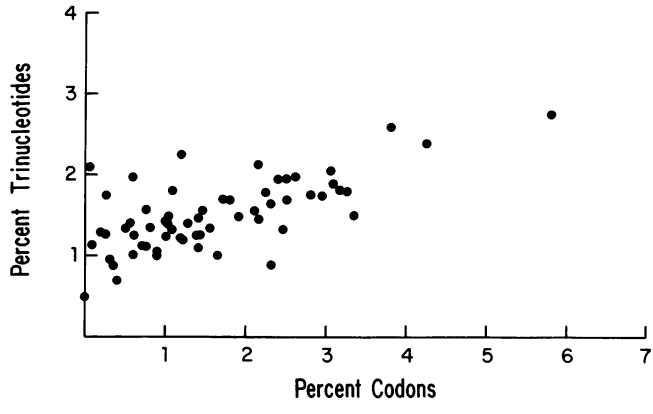

Figure 2. Linear regression analysis of codon usage on trinucleotide frequencies.

Codon usage was plotted against trinucleotide frequency in linear regression analysis. The coefficient of correlation was 0.649 (a coefficient of 1 indicates a linear relationship).

indicating that they occurred largely in the coding strand (Table 2). Conversely, 7 of the 10 least abundant trinucleotides had complement ratios less than one indicating a preference for noncoding DNA. These results were not necessarily unexpected given that codon usage represents one-third of the overall trinucleotide frequencies.

Other constraints acting on *E. coli* DNA have also been identified or proposed. Nearest neighbor biases in codon usage correlate with the level of gene expressivity (10). Our sample included genes that are expressed at

Table 2. Frequency (%) and Complement Ratios of the 10 Most- and 10 Least Abundant Trinucleotides

Rank	Trinucleotide	%	Complement Ratio[a]	Rank	Trinucleotide	%	Complement Ratio[a]
1	CTG	2.75	1.50	55	GGG	1.12	1.27
2	AAA	2.61	1.54	56	TAC	1.10	0.81
3	GCG	2.48	1.15	57	CTT	1.07	0.59
4	GAA	2.36	1.58	58	CTC	1.02	1.03
5	TGG	2.26	1.67	59	GAG	0.98	0.97
6	CGC	2.15	0.87	60	AGT	0.97	0.79
7	TGA	2.10	1.50	61	CCC	0.88	0.79
8	GGC	2.04	1.21	62	ATA	0.88	0.66
9	AAC	1.98	1.12	63	CTA	0.72	1.44
10	TGC	1.98	1.03	64	TAG	0.50	0.69

[a]ratio of the frequencies of trinucleotide to its complement.

high and low levels. Many prokaryotic DNAs also have an excess of RNY codons which may be a vestige of a primitive genetic code (11).

Codon Usage and Ordered Abundance of Oligonucleotides

The actual frequencies of longer oligonucleotides in E. coli DNA are unambiguously correlated with codon usage as can be seen by analysis of the ranking of individual oligonucleotides by ordered abundance curves in the accompanying paper (5). For example, the 3 most widely used codons (CTG, GAA and AAA) are the highest ranking in the trinucleotide ordered abundance curve (5) whereas 3 of the 4 least used codons (TAG, AGG, and AGA) ranked 64th, 53rd, and 45th in abundance. (TGA is exceptional; it ranks 63rd in codon usage but 7th in abundance). This also holds for tetra- through hexamers in that the most frequently used codons are found in oligomers in the upper end of the curves and vice versa for the least used codons. Examples of this relationship for tetra- through hexanucleotides are given in Table 3 where three of the most and four of the least frequently used codons are listed along with the ranking interval in which half of the oligomers containing it are found. For example, CTG is the most frequent codon and ranks first in trinucleotide abundance. Half of the 8 tetramers containing CTG (NCTG and CTGN) were found in the 1st through 11th most

Table 3. Correlation of most and least frequently used codons with tetra- through hexanucleotide ranking.

	rank	Frequency as: codon (%)	trinucleotide (%)	Ranking interval for 50% occurrance: tetra[a]	penta[b]	hexa[c]
CTG	1	5.78	2.75	1-11	1-94	1-859
GAA	2	4.26	2.36	5-17	5-135	5-903
AAA	3	3.83	2.61	3-14	7-103	5-697
AGA	61	0.20	1.31	150-256	706-1017	2756-4096
AGG	62	0.09	1.15	214-254	792-1022	2899-4089
TGA	63	0.07	2.10	5-44	46-294	N.D.
TAG	64	0.01	0.50	253-256	996-1024	3773-4096

[a] ranking interval from ordered abundance lists in which 4 out of 8 possible tetramers containing the codon were found among 256 tetranucleotides.

[b] ranking interval from ordered abundance lists in which 24 out of 48 possible pentamers containing the codon were found among 1,024 pentanucleotides.

[c] ranking interval from ordered abundance lists in which 128 out of 256 possible hexamers containing the codon were found among 4,096 hexanucleotides.

Table 4: Underrepresented tetranucleotides.

	frequency (%)[a]			number per[b]	complement
	ob	ex	(res)	74,444 bases	ratio[c]
CTAG	0.04	0.08	(-5)	27	palindrome
TATA	0.15	0.20	(-3)	114	palindrome
GAGG	0.17	0.22	(-3)	127	.90
GTCC	0.18	0.31	(-4)	137	.97
TAAG	0.22	0.31	(-4)	168	.96
CAAG	0.23	0.36	(-6)	175	1.08
GCTC[d]	0.27	0.33	(-3)	199	.73
TTGG[d]	0.27	0.44	(-7)	201	1.00
CCAA[d]	0.27	0.36	(-4)	201	1.00
TGTG	0.27	0.31	(-3)	204	1.15
ACAG	0.28	0.39	(-5)	207	.68
GGCC	0.31	0.44	(-5)	229	palindrome
GAAT	0.34	0.44	(-4)	253	1.05
GGCT	0.35	0.46	(-4)	261	.98
TGCA	0.40	0.48	(-3)	296	palindrome
GATC[d]	0.41	0.52	(-4)	309	palindrome
GGTT[d]	0.45	0.55	(-4)	340	.99
AACC[d]	0.46	0.56	(-4)	345	1.01
CGCG	0.57	0.67	(-3)	423	palindrome
CTGA	0.65	0.76	(-3)	485	1.41

[a] observed frequency (ob) and that expected (ex) by a 2nd order Markov chain in 74,444 bases.
[b] actual number in the 74,444 bases data set.
[c] ratio of each tetranucleotide frequency to its complementary tetranucleotide in coding DNA; palindromes have a value of 1.0.
[d] tetranucleotides that have a complementary sequence in the underrepresented list.

abundant tetranucleotides (or a ranking interval of 1-11); half of the 48 CTG-containing pentamers (e.g., NNCTG, NCTGN and CTGNN) were found by the 94th pentanucleotide of 1024 pentamers; and half of the 256 hexamers containing CTG (NNNCTG, NNCTGN, NCTGNN and CTGNNN) by the 859th hexanucleotide of 4096 hexamers. Hence, CTG oligonucleotides are found predominantly in the upper range of abundance curves and a similar conclusion holds for GAA and AAA, the next two most frequently occurring codons. Furthermore, infrequently utilized codons (TAG, AGG and AGA) were found in the lower end of the ordered abundance curves.

Nonpredicted Tetranucleotides

Although the foregoing indicates that overall frequencies of oligonucleotides is strongly correlated with codon usage, several tetra- through

hexanucleotides were over- or underabundant for reasons not correlated with codon usage. Among tetranucleotides, CTGG was the most abundant tetranucleotide occurring 726 times and CTAG the least abundant occurring 27 times out of 74,444 tetranucleotides; thus, there was a 27-fold variation between the highest and lowest abundant tetramer. Forty-five tetranucleotides were found that had residuals larger than ±3 by a 2nd order Markov chain (the probability of a residual greater or less than 3 is 10^{-3}).

There was an 18-fold variation in the frequency of tetranucleotides in the underabundant set: CTAG occurred only 27 times, CTGA occurred 485 times (Table 4). Of the 20 underabundant tetranucleotides, six were palindromes with equal complement ratios and one was the Dam methylation site (GATC). On a random basis, one-sixteenth of 20 tetranucleotides (or 1.25) would be expected to be palindromes. Hence, the underabundant set is significantly enriched for palindromic sequences. By contrast, the underabundant set was deficient in complementary tetranucleotides. Only 4 of the expected 10 underabundant tetranucleotides were complements of each other (TTGG with CCAA and GGTT with AACC) and occurred with equal frequency in both coding and noncoding DNA as seen by their complement ratios. Similarly, 17 of the underabundant set occurred at nearly equal frequency in both coding and noncoding DNA and have not, therefore, been selected for underabundance for coding reasons with respect to their complements. GCTC and ACAG had complement ratios substantially less than one and were more prevalent in the noncoding strand. CTGA had a ratio much greater than one and was enriched in coding sequences most likely for codon usage because CTG is the most widely used codon in E. coli (12).

These results contrasted with the 25 overabundant tetranucleotides of which none were palindromes (Table 5). Nearly half of the expected level were complements of each other, but none had complement ratios near unity. The 5 most abundant tetranucleotides (CTGG, GGCG, GAAG, GATG and GGTG) began with a trinucleotide that occurred more frequently in-frame (CTG, GGC, GAA, GAT and GGT) and ended with guanine which is most frequent in the first codon position (Table 1). For the remaining overabundant tetranucleotides, there was otherwise no obvious correlation with unusual codon usage patterns nor with complement ratios.

It was anticipated that underabundant tetranucleotides found by the Markov chain would be related at the trinucleotide level to overabundant tetranucleotides to compensate for codon usage requirements. For example, CTAG occurred only 27 times in 74,444 bases. To ensure sufficient CTA and

Table 5. Overrepresented tetranucleotides

	frequency (%)[a]			number per[b]	complement
	ob	ex	(res)	74,444 bases	ratio[c]
CTGG[d]	0.97	0.83	(+4)	726	1.91
GGCG[d]	0.81	0.65	(+5)	609	1.49
GAAG	0.72	0.56	(+6)	541	1.31
GATG[d]	0.64	0.52	(+4)	476	1.30
GGTG[d]	0.61	0.53	(+3)	455	1.41
TGCT	0.55	0.45	(+3)	411	1.21
CGCC[d]	0.55	0.46	(+3)	409	.67
CCAG[d]	0.51	0.42	(+3)	381	.53
CAGG	0.51	0.41	(+4)	380	.89
TATC	0.49	0.41	(+3)	369	1.73
CATC[d]	0.49	0.41	(+3)	363	.76
CAAC	0.49	0.39	(+4)	363	1.00
TGTT	0.47	0.38	(+4)	355	.94
TTGC	0.46	0.38	(+3)	345	.85
CGAC	0.45	0.36	(+3)	336	1.09
CACC[d]	0.43	0.33	(+4)	323	.71
ACAA[d]	0.40	0.33	(+3)	297	1.29
TTCC	0.40	0.32	(+3)	297	.85
TAAT	0.31	0.25	(+3)	234	.91
TTGT[d]	0.31	0.25	(+3)	230	.77
AATA	0.28	0.22	(+3)	213	.83
AGAG	0.26	0.20	(+3)	196	1.03
CTAC[d]	0.25	0.18	(+3)	186	1.10
GGAG	0.25	0.19	(+3)	185	1.02
GTAG[d]	0.23	0.16	(+4)	169	.91

[a] observed (ob) or expected (ex) frequency by a 2nd order Markov chain in 74,444 bases.
[b] actual number in the 74,444 bases.
[c] ratio of each tetranucleotide to its complementary tetranucleotide frequency in coding DNA.
[d] tetranucleotide having a complementary tetranucleotide in the over abundant set.

TAG for coding sequences, other CTA- and TAG-containing tetranucleotides must be overabundant, and two tetramers in the overabundant set of tetranucleotides contained CTA and TAG. The Dam methylation site GATC also appeared to be correlated with three overabundant tetranucleotides containing either GAT or ATC. These numbers do not, however, deviate significantly from expected number of 1.5. Moreover, the probability of finding 0, 1, 2, 3 or 4 tetranucleotides in a randomly chosen sample of 25 having a

trinucleotide in common with any tetramer in the underabundant list is 0.20, 0.33, 0.27, 0.21 and 0.05, respectively. This is based on a binomial distribution both with and without replacement of the chosen trinucleotide from the sample.

From the foregoing, it is concluded that the 45 over- or underabundant tetranucleotides occur at levels not predicted exclusively by codon usage. Their frequencies, therefore, are subject to constraints that are not immediately apparent. The possibility that the constraints occur at the level of larger oligonucleotide sequences of which the tetranucleotides are components is largely ruled out as described below.

Nonpredicted Penta- and Hexanucleotides

For 1024 pentanucleotides, only 8 did not fit values predicted by a 3rd order Markov chain (5), and had residual values larger than ±2 (5): TTGGA (-3), TGCGA (-4), GTGGC (-3), GACTG (-3), GAACC (-4), AAACC (+3), GTCTG (+3), and TCCTC (+3). The most abundant pentamer was TGCTG at 0.341% frequency and the rarest GCTAG at 0.007%, or a 50-fold variation from highest to lowest pentanucleotide. Also, of the 56 least abundant pentanucleotides, 38 contained CTA or TAG and the last 8 were NCTAG or CTAGN, as expected given the rarity of CTAG. Over- and underabundant pentanucleotides have an abundance of TG and GA (4 each) both of which have strong codon position biases. No other pentamers, including five-base restriction/modification recognition sites, were underrepresented which stands in contrast to bacteriophage T7 DNA where several tetra- through hexameric restriction/modification recognition sites have been selected against apparently to avoid exclusion by the host (13-15). Furthermore, no over- and underrepresented pentamer found by Markov analysis was related to the over- and underabundant tetranucleotides suggesting that those pentamers were set by constraints occurring at a level higher than tetranucleotides. The majority of pentamers containing the over- and underabundant tetranucleotides (Tables 4 and 5) also occurred at predicted values suggesting that selection for or against these tetramers did not take place at the pentamer or higher level.

4,001 hexanucleotides were well predicted by a Markov chain, having residuals ranging from 0 to ±1. Two hexamers with residuals of -5 are G+C palindromes: GCCGGC and GGCGCC occurring 5 and 6 times in 74,444 bases. By contrast, the six other G+C palindromic hexamers (CCCGGG, CCGCGG, CGCGCG, CGGCCG, GCGCGC, and GGGCCC) occurred at predicted levels but one of them (GGGCCC) was found only 4 times in the data set while the others ranged

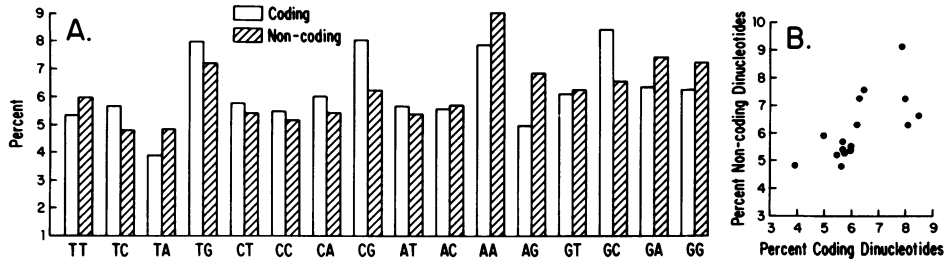

Figure 3. Comparison of dinucleotide frequencies in coding sequences to ribosomal and transfer RNA genes.

In A, dinucleotide frequencies in protein coding sequences (open are plotted next to those in ribosomal and transfer RNA genes (hatched); they are also plotted in linear regression analysis in B. The coefficient of correlation was 0.607 (a coefficient of 1.0 indicates a linear relationship).

from 11-37 occurrences. Thus, their actual levels varied considerably. It should be noted that GGCC and CGCG are components of 5 of these hexanucleotides and are underrepresented at the tetranucleotide level (Table 4). Also, sample size was limiting for the rare hexamers because eleven were not found in the data set and 24 were found only once; CTAG was the predominant member of these 35 hexamers. At the upper end, of abundance (5) CTG and CTGG dominated the highly abundant hexamers in which CTGCTG was most abundant, occurring 105 times.

On the Source of Sequence Variation in the E. coli Genome

The Markov chain rule did not accurately estimate the frequencies of di- and trinucleotides. However, it became relatively accurate at the tetranucleotide level and higher. This implies that constraints affecting nucleotide frequencies occur at the trinucleotide level or lower (16, 17). Given the wide variation in dinucleotide levels, we asked whether dinucleotide frequencies were set by a well-established and highly evolved translation mechanism which was pre-committed to a particular distribution of dinucleotides in codons, or did dinucleotide frequencies influence codon usage?

To approach this question, we reasoned that if codon usage had set dinucleotide levels, then sequences of genes that are transcribed, but not translated should have diverged free of any constraints imposed by codon usage. Ribosomal and transfer RNA's (12,336 bases) were analyzed and compared to coding sequences (47,358 bases); 5' and 3' flanking sequences were not analyzed because of the likelihood of encountering amino acid

coding sequences immediately upstream from a promoter or downstream from a terminator. Figure 3A compares dinucleotide frequencies from each data set, which have also been plotted against each other in Figure 3B. A straightline regression analysis gave a coefficient of correlation of 0.607, indicating that the values are more similar than dissimilar. In both data sets, AA was among the most abundant and TA the least frequent. Nearly identical frequencies were found for seven dinucleotides (i.e., GT, CA, AC, TC, CT, AT, TT, and CC). The largest differences occurred in the frequencies of GC, CG, GG, GA and AG. Ribosomal and transfer RNA genes are abundant in purine-purine dinucleotides AA (9.1%), GA (7.5%), GG (7.25%), and AG (6.9%), perhaps because of the high degree of secondary structure in rRNA. Also, C-G base pairs are thermodynamically more stable than A-T base pairs, so the high level of GG and CG may reflect their ability to stabilize intrastrand helices in structural RNAs.

Despite functional differences between the coding and noncoding sequences, the overall pattern of dinucleotide frequencies in the two samples were remarkably similar, implying that codon usage was not the sole driving force in setting dinucleotide frequencies (18). However, most E. coli DNA codes for proteins, so that the transcription and replication machinery may have evolved to function most efficiently on coding sequences and secondarily imposed limits to divergence of sequences that do not code for polypeptides.

Concluding Comments

In this report, we have provided estimates of the influence of codon usage on setting oligonucleotide frequencies in the E. coli genome. Not unexpectedly, trinucleotide frequencies were strongly correlated with codon usage by linear regression analysis. Analysis of ordered abundance data also showed that tetra- through hexanucleotides were also strongly correlated with codon usage. Nonetheless, not all sequences could be explained by codon usage. Hence, 45 tetranucleotides and 8 penta- and 6 hexanucleotides were detected by Markov chain analysis that were over- or underabundant by virtue of their large, absolute residual values. Whether or not these sequences underwent selection to set and to maintain their levels is unknown. None were obvious components of known regulatory sequences associated with structure of DNA and its expression (i.e., promoters/terminators, etc.). Nonetheless, six of the tetra- and two of the hexanucleotides were palindromes and fell into the underabundant group. Because the frequency of palindromes in the group is far in excess of what would be

expected by chance alone, their levels appear to have undergone negative selection. What the nature of the selection might be and at what level it occurs (structure and function DNA or an underlying molecular process, e.g., restriction and modification) remains to be shown.

ACKNOWLEDGMENTS

The authors thank Chris Williams for programming the regression analysis, and Suzette Lay for her expert typing of the manuscript, especially the tables. This work was supported largely by a U.S. Army Research Office Training grant (D-AAG29-83-G-0111) to J. A. and partially by an NIH grant (CA-34066) to R.I.

REFERENCES

1. Wyman, A.R., Wolfe, L.B., and Botstein, D. (1985) Proc. Natl. Acad. Sci. U.S.A. 82, 2880-2884.
2. Leach, D.R.F., and Stahl, F. (1983) Nature (London) 305, 448-451.
3. Trifonov, E.N., Konopka, A.K., and Jovin, T.M. (1985) FEBS 185, 197-202.
4. Peck, L.J., and Wang, J.C. (1985) Cell 40, 129-137.
5. Phillips, G.J., Arnold, J., and Ivarie, R. (1987) Nucleic Acids Res., Accepted for publication.
6. Arnold, J., Eckenrode, V.K., Lemke, K., Phillips, G.J., and Schaeffer, S.W. (1986) Nucleic Acids Res. 14, 239-254.
7. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
8. Ikemura, T. (1981) J. Mol. Biol. 151, 389-409.
9. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
10. Yarus, M., and Folley, L.S. (1985) J. Mol. Biol. 182, 529-540.
11. Shepherd, J.C.W. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 1596-1600.
12. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981) Nucleic Acids Res. 9, r43-74.
13. McClelland, M. (1985) J. Mol. Evol. 21, 317-322.
14. Sharp, P.M. (1986) Mol. Biol. Evol. 3, 75-83.
15. Sharp, P.M., Rogers, M.S., and McConnell, D.J. (1985) J. Mol. Evol. 21, 150-160.
16. Nussinov, R. (1984) J. Mol. Evol. 20, 111-119.
17. Lipman, D.J. and Wilbur, W.J. (1983) J. Mol. Biol. 163, 363-376.
18. Nussinov, R. (1981) J. Mol. Evol. 17, 237-244.